HUF 2016
HLRS Data Management Evolution

Björn Schembera (schembera@hlrs.de)

New York City

September 1st 2016

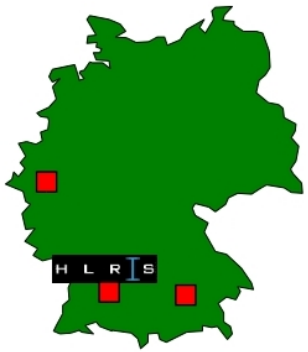
Outline

HLRS

Höchstleistungsrechenzentrum Stuttgart

Overview: HLRS

- Part of the University of Stuttgart
- One of 3 German National Centers (GCS)
- $>$ 90 staff
- Operation of HPC systems
- Focus: Engineering / Global System Science
- Research: Programming Models, Visualization, ...
- Teaching / Training: New building in late 2016

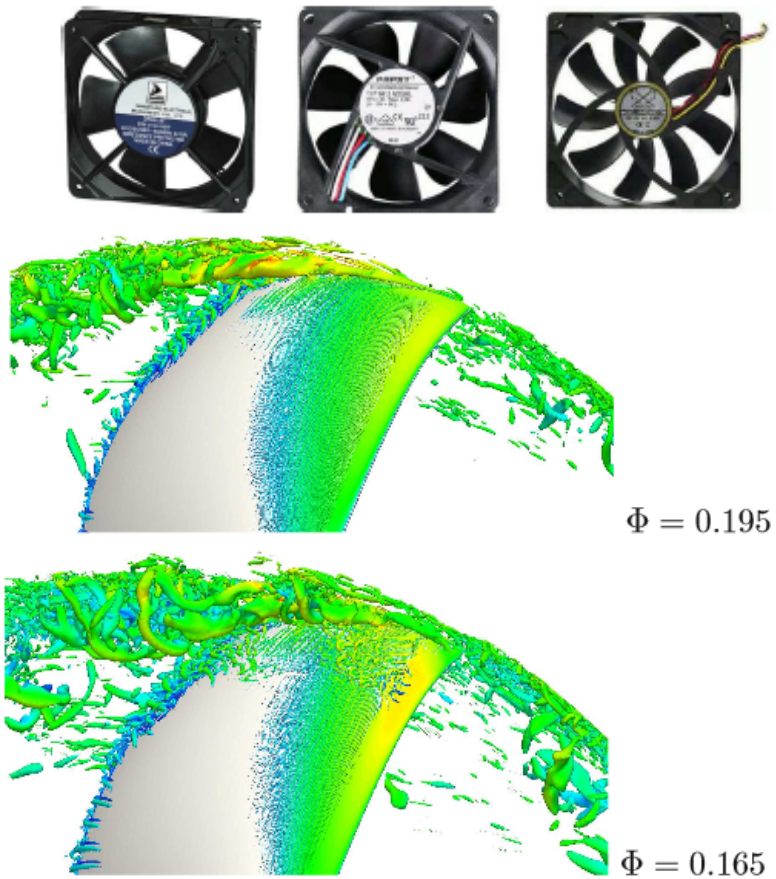


Overview: Projects

Main areas of users research:
Aeroacoustics, Aerodynamics, Astrophysics, Bioinformatics, Combustion, Fluid-Structure Interaction, Meterology, Medical Imageing, Nanotechnology, Solid State Physics, Turbulence Phenomena
Example:

- Dr. Meinke (RWTH Aachen): flow around axial fan [MAF]
- 1 billion cell mesh
- 100 TB of result data
- Statistical analysis
- New methods detect structures within turbulence
- 1PB data sets in the future

- Axial fans generate annoying noise



Systems

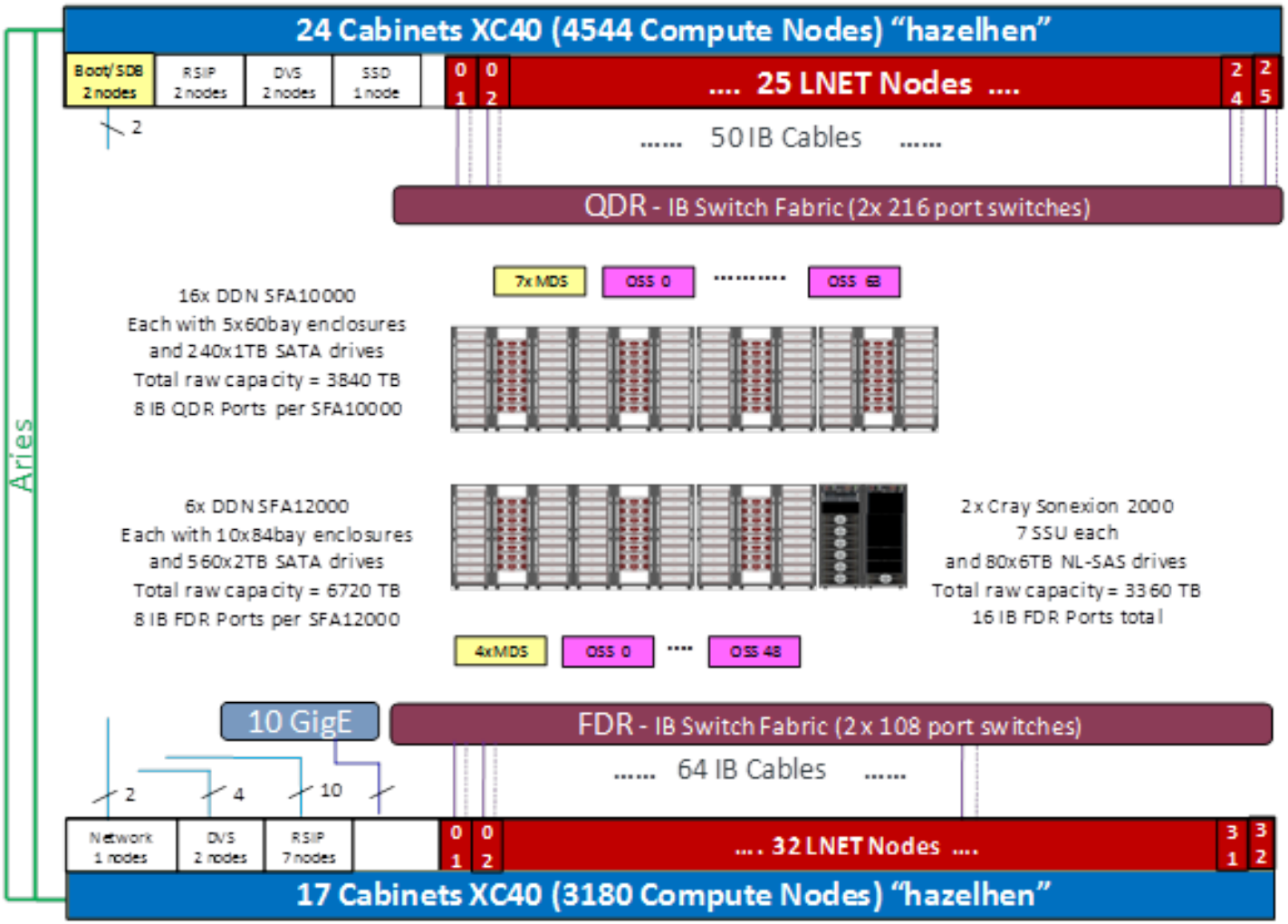
CRAY XC40 „Hazelhen“

(Hermit < Hornet < Hazelhen < H..... ?)

CRAY XC40 „Hazelhen“

- Hazelhen = Hornet (2014) + upgrade (2015)
- Main Production system since Oct 2014
- 7712 nodes, each 2 sockets w/ 12 cores Intel/Haswell@2.5GHz and 128GB main memory = 185088 cores and 1PB of main memory
- Interconnect: CRAY Aries
- Performance 7.42PF/s peak / 5.64PF/s Linpack / 136TF/s HPCG
- Power consumption 3.2MW

CRAY XC40 „Hazelhen“ Conceptual Architecture



NEC machines

SX-ACE

- 64 nodes w/ 4 core vector CPU
- Interconnect: NEC IXS 8 GB/s
- 250TB NEC ScaTeFS filesystem
- Performance 16TF/s
- 30kW

LAKI cluster

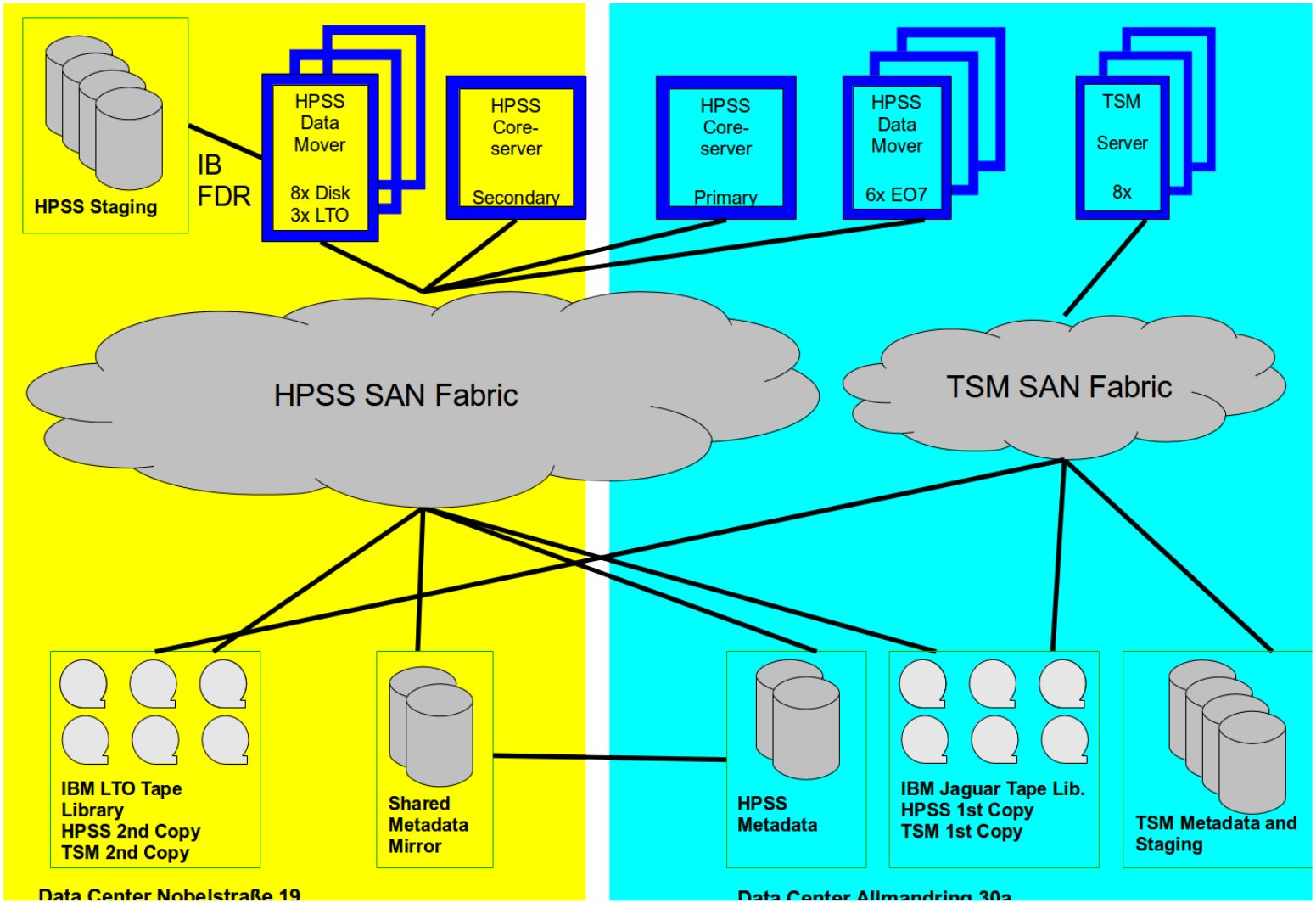
- 2 clusters
- 751 nodes of different types (SandyBridge, Nehalem, Interlagos)
- Interconnect: Infiniband, 10GE
- 350TB Lustre filesystem
- Performance 154TF/s + 74TF/s

HPSS

HPSS: Overview and Hardware

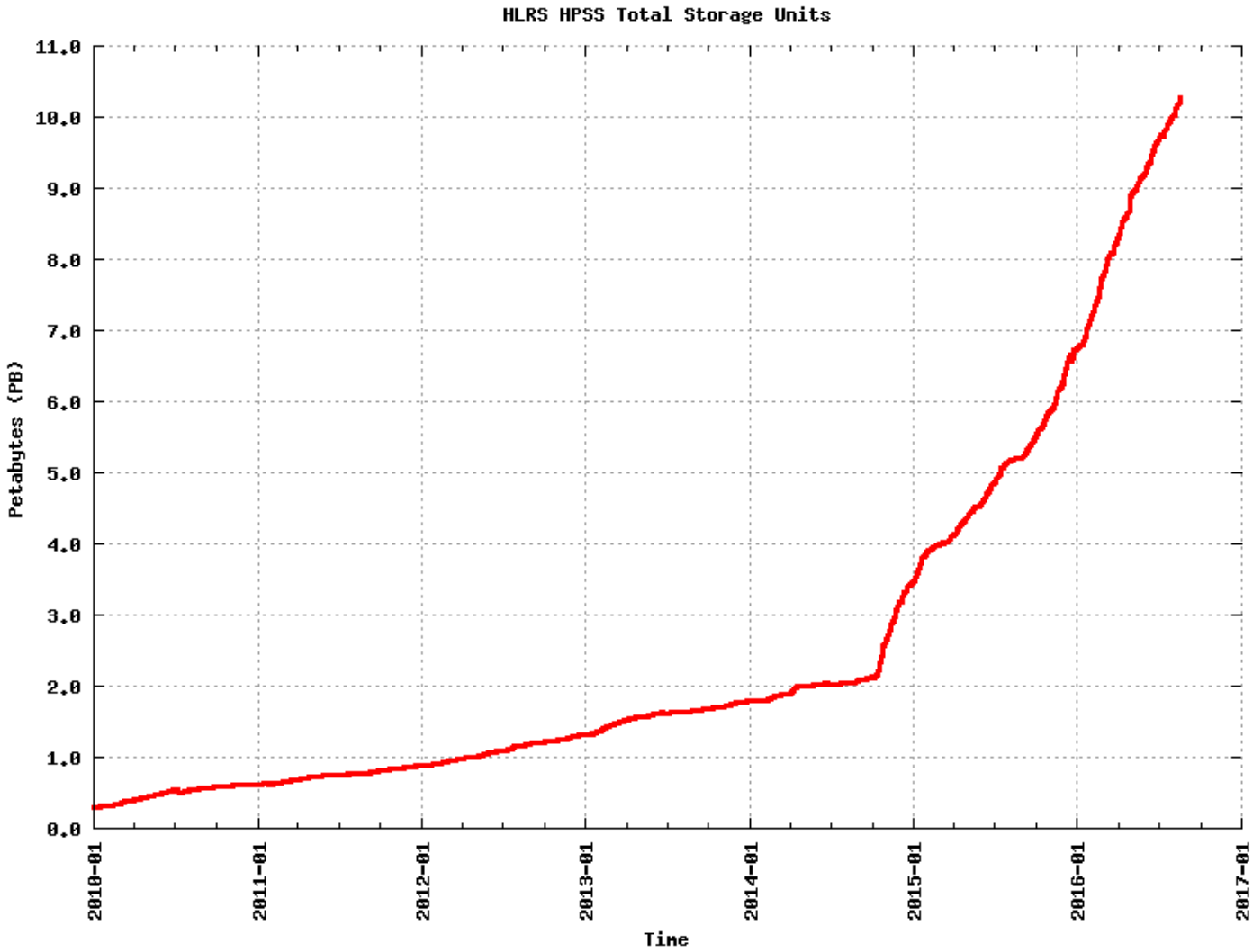
- Major technology refresh in 2014
- Libraries
  - Shared use of libraries with TSM
  - Spatially separated 1st/2nd copy
  - 1st: IBM TS3500 15 frames w/ 24 EO7 + 8 E05
  - 2nd: IBM TS3500 10 frames w/ 32 LTO6 drives
- Core Servers: 1 (+1 standby) (IBM x3655 M4)
  - 2x8 cores, 128GB, 10GE, 4x8GB FC
  - RHEL 6.4
  - HPSS 7.4.2p1
  - DB2 v10.5
- Data Movers: 17 (IBM x3655 M4)
  - 2x8 cores, 64GB, 10GE, 4x8GB FC
  - 8 disk movers, 6 E07 movers, 3 LTO6 movers
- Metadata Storage: 1+1 (IBM Storwize V3700)
  - 6TB Storage
  - SSD for tablespaces and logs
- Disk 4x NEC SNA460 + Disc Enclosure SNA060
  - 512TB
  - InfiniBand FDR Connection
  - each 5GB/s (with HPSS)
- User interfaces: PFTP, VFS, GridFTP@VFS

Upgrade of HPSS: Conceptual Architecture



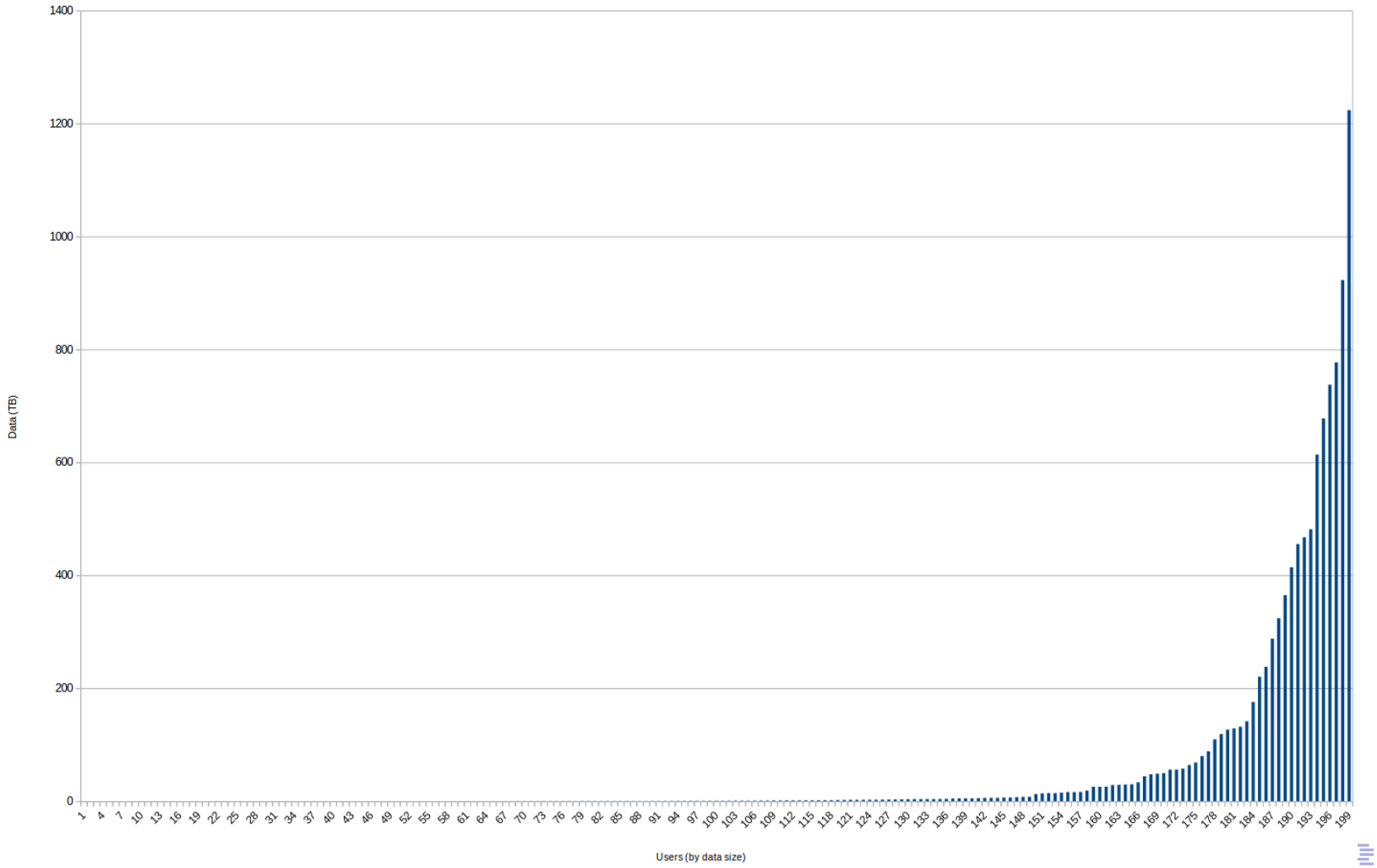
HPSS: Total Utilization (as of Aug 18th 2016)

Total Name Space Objects: 8.461.624



HPSS: Users / Data per user distribution / CoS

Active users: approx. 200 (Per User AVG: 52TB / MEDIAN: 0.64TB)
4 CoS: All w/ 2nd copy on LTO, one striped across 4 tapes on 1st copy



HPSS: Repacking

- 2014: LTO6 for 2nd copy
- Repacking of 2800 EO5 tapes and approx. 500 E07 to LTO6
- Started 1/2015
- Done 90%
- Issues like orphaned segments, read errors

HPSS: Future Work

- Lustre-HSM
  - Test installation
  - Installed agent node with CEA HPSS copytool
  - Next Step: Robinhood
  - Q3/2016
- Local File Transfer Movers
  - Attach Lustre directly to the movers
  - Testing planned for Q4/2016

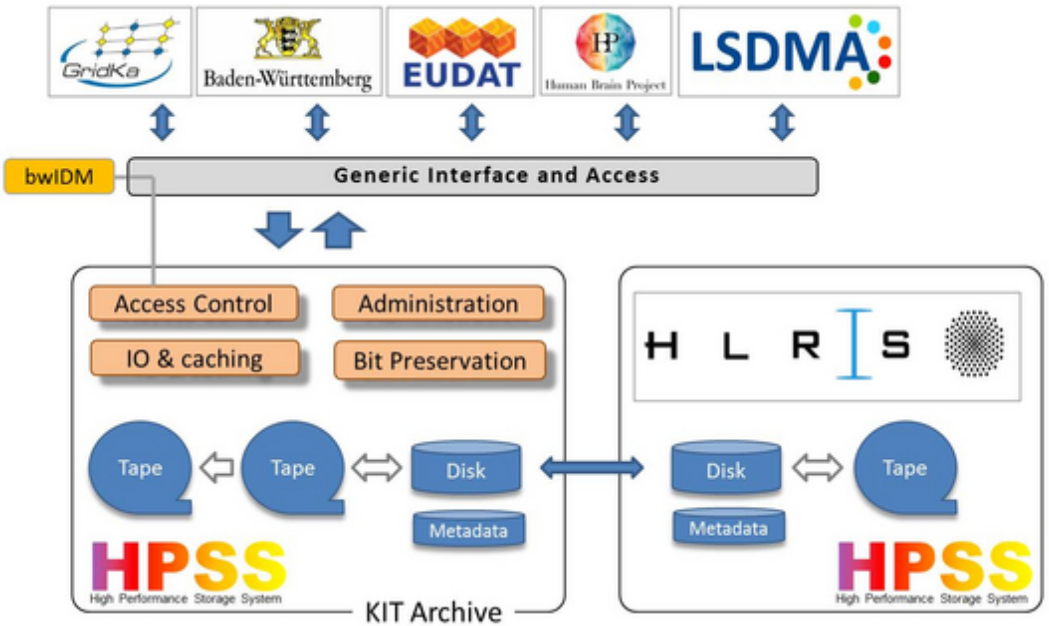
From LEIBOVICI Thomas <thomas.leibovici@cea.fr>
Subject **Re: LustreHSM/HPSS**
To Björn Schembera <schembera@hlrs.de>
Cc lustre-hsm@cea.fr <lustre-hsm@cea.fr>

```
Hi Björn,

On 08/25/16 11:56, Björn Schembera wrote:
| archive_id = 666

* This value looks suspicious to me '^^'

A few words about this parameter:

Lustre-HSM can manage several storage backend (AFAIK, up to 32). Each one
You can select the backend where you want to archive by specifying "--arch

    lfs hsm_archive --archive 1 file

You can control the default archive by changing the value of:

# cat /proc/fs/lustre/mdt/lustre-MDT0000/hsm/default_archive_id
1

If you just have 1 archive backend, just set archive_id = 1 in copytool cc

As you specify 666 > 32, it returns "invalid argument".

* This value also looks wrong :  "subsys_id   = 666 "

It should be your HPSS susbsys index (usually 1).

Conclusion: stop worshipping Hell and specify reasonable parameters 😀

* I also suggest you turn it on: "clean_non_ascii".
HPSS may not like to have filenames with special characters...

Best Regards,
Thomas
```

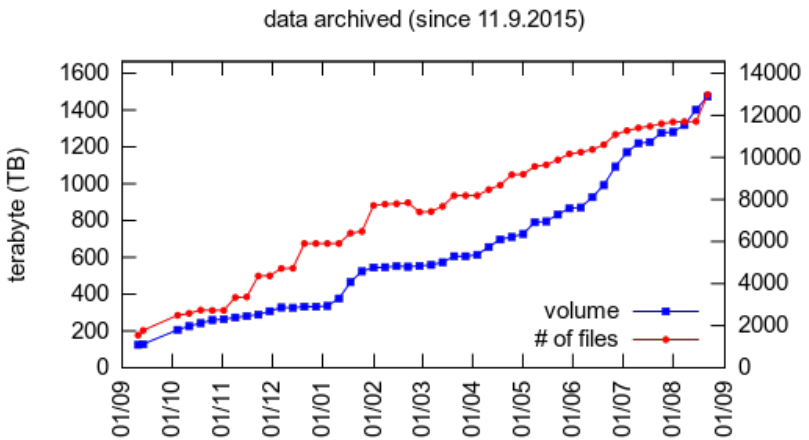
bwDataArchiv

bwDataArchiv

- Research project with Karlsruhe Institute of Technology (KIT) [BWDA]
- Long term storage for research data
- HLRS acts as project partner and „first client“
- Selected HLRS users already transfer data from HLRS to KIT (80km/50mi)

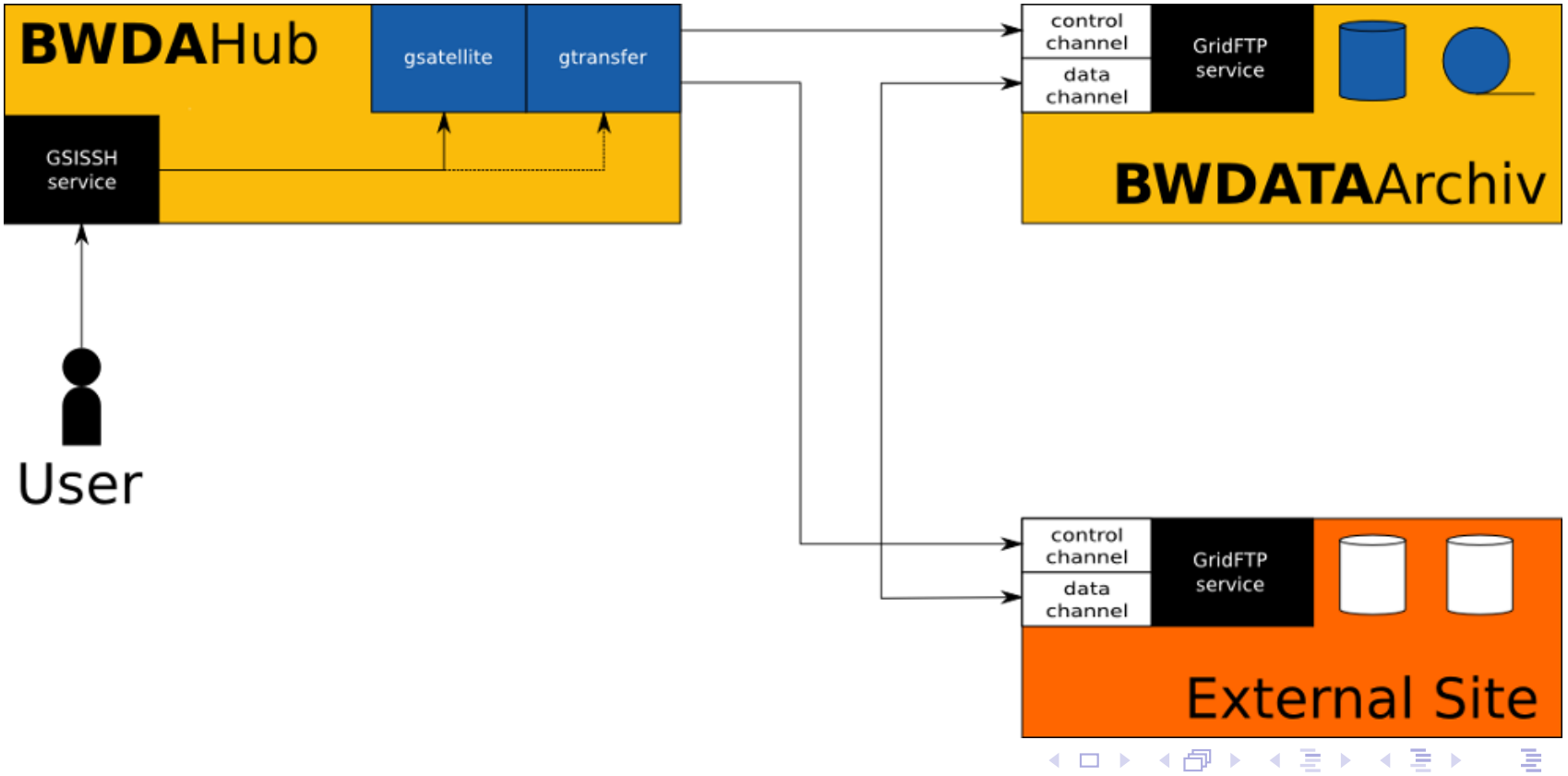


BWDATA ARCHIV



bwDataArchiv

- Dedicated node *bwdahub* for organizing transfers
- gtransfer as wrapper tool for GridFTP based data transfer [GT]:
  simplifies use; host aliases; transfer continuation; optimized performance
- gsatellite for scheduling data transfers [GS]
- GridFTP on top of VFS for HPSS@HLRS access



Summary

Summary

- Cray XC40 in production since 2014
- Upgrade of HPSS infrastructure completed in late 2014
- Data grows exponentially (as expected)
- Evaluating new technologies: LustreHSM and LFT Movers
- bwDataArchiv: Long-term archiving

Thank you for your attention!

References

[BWDA] https://www.rda.kit.edu/
[GS] http://bitly.com/2bg4Fry
[GT] http://bit.ly/gtransfer
[HLRS] http://www.hlrs.de
[MAF] https://www.hlrs.de/en/solutions-services/customer-projects/prediction-of-the-turbulent-flow-field-around-a-ducted-axial-fan/